Pedestrian Recognition using Cross-Modality Learning in Convolutional Neural Networks

Danut Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, Abdelaziz Bensrhair

Pedestrian Recognition using Cross-Modality Learning in Convolutional Neural Networks

Dănuţ Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, and Abdelaziz Bensrhair

***Abstract*—The combination of multi-modal image fusion schemes with deep learning classification methods, and particularly with Convolutional Neural Networks (CNNs) has achieved remarkable performances in the pedestrian detection field. The late fusion scheme has significantly enhanced the performance of the pedestrian recognition task. In this paper, the late fusion scheme connected with CNN learning is deeply investigated for pedestrian recognition based on the Daimler stereo vision dataset. Thus, an independent CNN for each imaging modality (Intensity, Depth, and Optical Flow) is used before the fusion of the CNN's probabilistic output scores with a Multi-Layer Perceptron which provides the recognition decision. We propose four different learning patterns based on Cross-Modality deep learning of Convolutional Neural Networks: (1) a Particular Cross-Modality Learning; (2) a Separate Cross-Modality Learning; (3) a Correlated Cross-Modality Learning and (4) an Incremental Cross-Modality Learning model. Moreover, we also design a new CNN architecture, called LeNet+, which improves the classification performance not only for each modality classifier, but also for the multi-modality late-fusion scheme. Finally, we propose to learn the LeNet+ model with the incremental cross-modality approach using optimal learning settings, obtained with a K-fold Cross Validation pattern. This method outperforms the state-of-the-art classifier provided with Daimler datasets on both non-occluded and partially-occluded pedestrian tasks.**

***Index Terms*—cross-modality learning, deep learning, multi-modal images, late-fusion.**

I. Introduction

THE ability to detect and classify objects is the fundamental requirement for designing intelligent application systems, like autonomous vehicles, and driver assistance systems. Pedestrian detection is one of the main concerns of transport safety and security, where an optimal Advanced Driver Assistance System (ADAS) for pedestrian detection is essential to reduce the number of traffic accidents and life-threatening injuries. The most frequent traffic accidents are caused by driving errors due to fatigue, discomfort, or using the phone while driving, but also by pedestrians' illegal and/or unsafe behavior. These accidents could be dramatically reduced if such human errors could be entirely eliminated.

Dănuţ Ovidiu Pop is a PhD student in the RITS Team, INRIA Paris, 2 Rue Simone IFF, 75012 Paris, France in collaboration with Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France and Department of Computer Science, Babeş-Bolyai University, 7- 9 Universitatii street, 400084 Cluj-Napoca, Romania. email: danut-ovidiu.pop@inria.fr

Dr. Alexandrina Rogozan is an Associate Professor at Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France. email: alexandrina.rogozan@insa-rouen.fr

Dr. Fawzi Nashashibi is the head of the RITS Team at INRIA Paris, 2 Rue Simone IFF, 75012 Paris, France. email: fawzi.nashashibi@inria.fr

Dr. Abdelaziz Bensrhair is a Professor at Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France. email: abdelaziz.bensrhair@insa-rouen.fr

In order to prevent collisions between vehicles and pedestrians or obstacles, ADAS systems generally use multi-sensor systems and/or a camera network to gather road traffic information; then the modality-specific processing components extract relevant features that are inserted into recognition components to be classified. If a potentially risky situation is detected, these systems either provide a warning to the driver and/or to the pedestrians, or even apply the brakes autonomously.

In recent years, a wide variety of studies have been carried out on ADAS systems. Since 2013, BMW cars have been equipped with a driver assistance package for pedestrian warning, based on infrared night-vision and monocular vision cameras. The Mercedes system has combined stereo vision cameras with long, medium and short-range radars to monitor the area in front of the vehicle. In 2016 the Continental company designed an advanced radar sensor (standard for VW Tiguan) able to distinguish both pedestrians and objects, at a distance of up to 170 meters. Since 2013, the Nissan company has developed a system which perceives the vehicle's environment, including the road, pedestrians, and other vehicles.

Nonetheless, the issue of ADAS systems remains an open challenge for researchers to develop more significant improvements because there still exist traffic situations that could dramatically compromise the safety of the road users concerned, especially the most vulnerable ones (i.e., pedestrians). Existing ADAS systems still do not provide a suitable result in such traffic situations, especially in crowded urban environments where the ADAS's progress is hindered by the difficulty of detecting all partially occluded pedestrians and the problem of running efficiently in extreme weather conditions. Furthermore, an ADAS system must ensure a complete and reliable real-time functionality. We deem it is also essential for the classification component of an ADAS system to distinguish the obstacle types (pedestrian, cyclist, child, old person) in order to adapt the system's behavior according to the estimated risk level.

A pedestrian detection system has three main components: the sensors used to capture the visual data, the modality- image processing components and the classification components. In general, all these components are processed and developed together to obtain a high detection performance, but sometimes each element could be investigated separately according to the target application. This paper is concerned with improving the classification task, which is the central part of the pedestrian detector, following the cross-modality learning methodology we proposed in [29]. We also explore a new Particular Cross-Modality Learning method within an original CNN classifier

architecture, which we called LeNet+.

In recent research studies, deep learning neural networks including Convolutional Neural Networks (CNNs) like AlexNet [18], GoogleNet [34], VGG [31] have usually led to improvements in classification performance, due to their capacity to learn discriminatory features from raw pixels. These CNNs differ in size and depth according to the objects that need to be classified. Thus, with an increase in the complexity of the classifier's problem, the CNN's size and depth also increase, which usually enhances the CNN's performance.

The drawback of CNNs with very large and complex architectures, such as GoogleNet, VGG, is that they require considerable computing power and a vast storage space, especially for the off-line learning process, but also to a lesser degree for the on-line classification applications. This problem has been partially solved since for the off-line step the CNNs could be learnt on an expensive powerful network of computers, but it could be an unsolved problem for several on-line embedded applications. Indeed, the CNNs involved in an ADAS system should fulfill some requirements to become a feasible solution for on-board implementation in a vehicle.

The question is: could we adapt a vast CNN architecture to be a feasible solution in order to upload it into a cheap embedded processing module or should we create a new one to fit on the required ADAS settings? Increasing the CNNs complexity (architecture and learning settings), the classifier models require higher computing power for off-line learning and on-line applications that lead to the purchase of more powerful and expensive GPUs. Therefore, we chose to propose a compact, but efficient CNN architecture for the pedestrian recognition task, well-suited to small-size multi-modal images derived from stereo vision.

Deep learning classification methods associated with multi-modality images and different fusion patterns have achieved notable performances. In this paper, we further investigate deep learning models proposed in [29], where two problems were explored: whether one modality could be exclusively used for training and validation of the classification model used to recognize pedestrians in another modality or together with other modalities to improve the classification model's learning in each modality. This paper investigates how a multi-modal system could be learnt when data in one of the modalities is scarce (i.e. many more images in the visual spectrum than depth). If the system is learnt on multi-modal data, could it still work when the data from one of the domains is missing? Could the learning process be improved if it uses a different image modality validation set than the training set? To the best of our knowledge, this issue, which we call cross-modality learning, has not yet been investigated for the pedestrian recognition task. This paper sets out to evaluate this cross-modality concept through various experiments based on the Daimler stereo vision dataset [10] and will allow us to chose the most promising one for this pedestrian classification task.

The main contribution of this paper is concerned about investigating different cross-modality learning approaches for deep neural networks aimed at the pedestrian recognition task (non-occluded and partially-occluded samples) using various sensor modalities. It also proposes a new variation of the LeNet architecture and provides results for a late-fusion approach.

The paper is organized as follows: Section 2 briefly shows our main contribution and some existing approaches from the literature. Section 3 presents an overview of our system, Section 4 presents the classification architectures and the associated learning methods based on Cross-Modality deep learning of CNNs. Section 5 describes the experiments and their results on the Daimler datasets. Finally, Section 6 presents our conclusion.

II. Related Work

The pedestrian detection issue has attracted considerable interest over the last decade, resulting in a wide variety of detection methods. Pedestrian detection approaches can generally be classified in two categories:

- handcrafted features models such as Integral Channel Features [6], Histograms of Oriented Gradients (HOG) [5], Local Binary Patterns (LBP), Scale Invariant Feature Transform [37], among others [30], [12], followed by a trainable classifier such as a Support Vector Machine (SVM) [12], Multi-Layer Perceptron (MLP), boosted classifiers [6] or random forests [7], [2];
- deep learning neural networks models, particularly Convolutional Neural Networks [15], [13], [1], like LeNet [20], AlexNet [18], GoogleNet [34], VGG [31] which have to extract implicit features for classification purposes.

A. Handcrafted Features Models

We chose to briefly present only the state-of-the-art models given with the Daimler datasets, since our cross-modality learning models are developed on those datasets. A mixture-of-experts (MoE) framework performed with HOG and LBP features, and MLP or linear SVM classifiers was presented in [10], [11].

In the HOG/linSVM MoE, the HOG descriptor was computed with 12 orientation bins and 6 x 6 pixel cells, accumulated for overlapping 12 x 12 pixel block with a spatial shift of 6 pixels, and then those features were inserted into liner SVM [10].

In the HOG+LBP/MLP MoE, the HOG and LBP features were inserted into MLP [11]. The HOG descriptor was applied with 9 orientation bins and 8 x 8 pixels cells, accumulated for overlapping 16 x 16 pixels blocks with a spatial shift of 8 pixels. The LBP descriptor was applied using 8 x 8 pixels cells and a maximum number of 0-1 transitions of 2. Those feature-based MoE models are learnt using a classical learning methodology where both training and validation were done on the same modality: Intensity, Depth or Optical Flow.

B. Deep Learning Neural Network Models

A deformation part-based model combined with a deep model based on a restricted Boltzmann Machine for pedestrian detection is presented in [22]. The deformation-part component receives the scores of pedestrian body-part detectors and

provides a decision hypothesis to the deep model in order to discriminate the visibility correlation among overlapping elements at multilayers. This approach was applied not only on the Daimler datasets but also on the Caltech, ETH and CUHK datasets. A deep unified model that conjointly learns feature extraction, deformation handling, occlusion handling and classification evaluated on the Caltech and ETH datasets for pedestrian detection was proposed in [23].

A solution for detecting pedestrians at different scales and evaluated on the Caltech data set by combining three CNNs was proposed in [9]. A cascade Aggregated Channel Features detector is used in [40] to create pedestrian candidate windows followed by a CNN-based classifier for assessment purposes on monocular Caltech and stereo ETH data sets.

Two CNN-based fusion methods (early and intermediate fusion architectures) of thermal and visible images were presented in [38] and evaluated on the KAIST pedestrian data set. The early fusion approach merges the information of these modalities at the pixel level, the intermediate fusion method generates a feature representation for each modality using separate sub-networks before classification. The authors concluded that intermediate fusion has greater classification accuracy than early fusion.

We presented an early fusion versus late fusion comparison on the non-occluded Daimler stereo vision dataset in [27]. The early fusion approach integrates three image modalities (Intensity, Depth and Optical Flow) by concatenating them to learn a single CNN. The late fusion approach consists in fusing the probabilistic output scores of three independent CNNs, trained on different image modalities (Intensity, Depth and Optical Flow) by an SVM classifier.

We concluded that the early-fusion approach is less efficient and robust than the late-fusion model. Moreover, the early-fusion model requires high image calibration and synchronization. The early-fusion training method is more constrainable since for a given image frame it needs an item for each modality, and therefore the classifier requires more samples to learn the problem. With the early-fusion model, it is impossible to take advantage of cross-dataset training methods, by using modality images from different unimodal and/or multi-modal datasets where all the modalities involved are not acquired and/or annotated. The early fusion method does not allow one to improve the learning by extending the number and the variety of items through the cross-modality learning we proposed in [29].

In the literature, for the late fusion architectures, the learning is performed independently on each modality, with annotated images provided exclusively from that modality. To the best of our knowledge, no study has been carried out on cross-modality learning for pedestrian recognition, but only on cross-dataset learning. In [17], the authors proposed an incremental cross-dataset learning algorithm for the pedestrian detection problem. A synthetic dataset (Virtual Pedestrian dataset [36]) is used for basic training and two distinct real-world datasets (KITTI Vision Benchmark Suite and the Daimler Mono Pedestrian Detection Benchmark) for fine-tuning the models and for evaluation.

III. Overview of our System

The goal of the work presented in this paper is to improve the late-fusion learning of pedestrian classifiers by using a cross-modality approach. We propose different learning methods based on Cross-Modality deep learning of CNNs:

- a Particular Cross-Modality learning method where a CNN is trained and validated on the same image modality, but tested on a different one;
- a Separate Cross-Modality learning method which uses a different image modality for training than for validation;
- a Correlated Cross-Modality learning method where a unique CNN is learnt (trained and validated) with Intensity, Depth and respectively Optical Flow images for each frame;
- an Incremental Cross-Modality learning where a CNN is learnt with the first images modality frames, then a second CNN, initialized by transfer learning on the first CNN, is learnt on the second image modality frames, and finally a third CNN initialized on the second CNN, is learnt on the last image modality frames.
- an improvement of the incremental cross-modality learning due to a new CNN architecture that we proposed together with K-fold Cross-Validation of both the learning rate and epoch numbers.

We examine all these methods with the classical learning one where each CNN is learnt and evaluated on the same image modality. We also compare all these learning patterns with the classical learning approaches within the MoE framework proposed in [10], [11] and deep Boltzmann-Machine [22] for the recognition of both partially occluded and non-occluded pedestrians.

The following section describes the architecture and the corresponding settings for each of the cross-modality learning methods.

IV. Proposed Architectures

This paper concerns fusing stereo-vision information between three modalities: Intensity (I), Depth (D) and Optical Flow (F). We investigate the late-fusion architecture using five distinct approaches for the learning of the CNN-based classifiers: a classical intra-modality approach and four different methods for the cross-modality approach. These methods differ in the manner in which they are used to train and valid the CNNs.

A. Classical Learning

The Classical Learning (CL) method involves that both training and validation of a model are done on the same image modality. For each image modality, a classifier model is fitted on the respective training dataset; successively, the fitted model is used to predict the labels for the observations in the validation dataset; and finally, the test dataset is used to provide an unbiased evaluation of the final model fitted on the learning dataset (union of training and validation datasets). For the classical learning approach, we have trained, validated and evaluated each CNN with the same imaging modality among Intensity, Depth, and Optical Flow (see Fig.1).

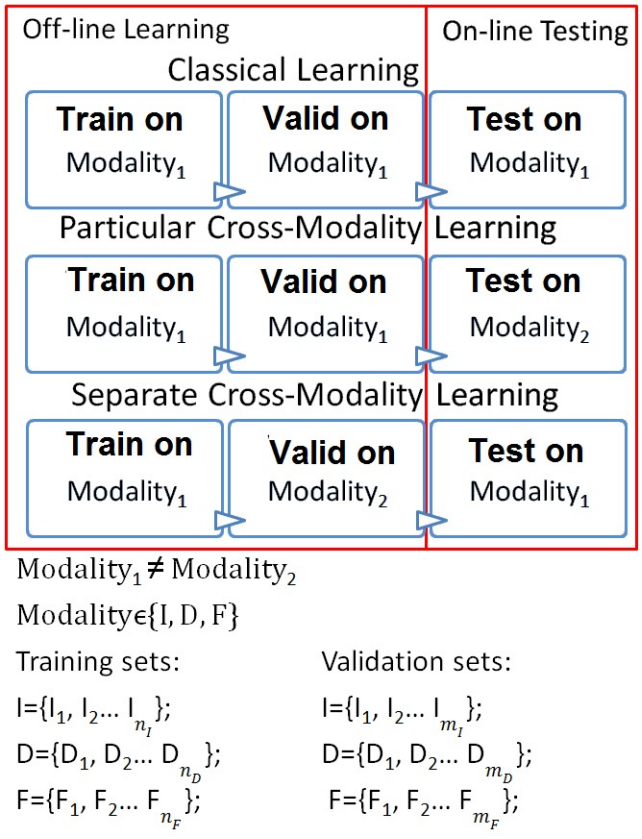


Fig. 1. The classical learning approach uses the same image modality for the training, validation, and testing processes. The Particular Cross-Modality learning uses the same image modality for training and validation, but a different one for testing. The Separate Cross-Modality learning uses the same image modality for training and testing, but a different one for validation.

B. Particular Cross-Modality Learning

Particular Cross-Modality Learning (PaCML) carries out the learning process on the same image modality, although the training and validation sets are disjoint, and the performance is evaluated on a different modality. This approach shows whether the automatic annotation of modality images could be extracted with a classifier trained with different modality data (see Fig.1).

C. Separate Cross-Modality Learning

Separate Cross-Modality Learning (SeCML) carries out the learning process when the modality of the training set differs from that of the validation set. The testing set belongs to the same modality as the training set (see Fig.1). This approach could improve the generalization power of CNN and shows how we could train a system when one of the imaging modalities is limited.

D. Correlated Cross-Modality Learning

The Correlated Cross-Modality Learning (CoCML) approach learns a single CNN, where the data training set consists of frames with distinct image modalities: Intensity I_i, Depth D_i and Flow F_i with i=$\overline{1,n}$ (see Fig. 2). The CNN model is validated in two different ways: on a multi-modal validation set (a stack of images from the same frames for different image modalities) and respectively on a uni-modal validation set. The training and validation sets are disjoint.

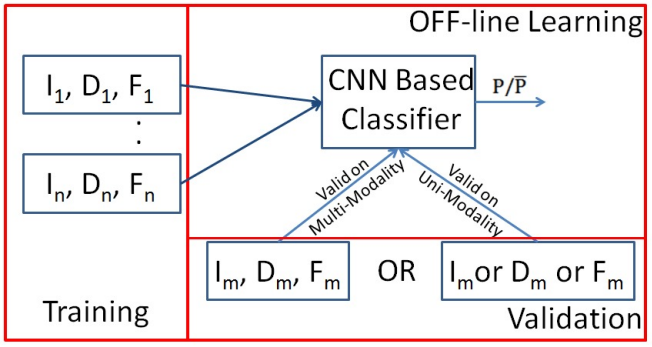


Fig. 2. The Correlated Cross-Modality Learning. The learning data consists in Multi-Modal Correlated images presented successively to the CNN for training and respectively in Multi-Modal or Uni-Modal images for validation. I$\in\{I_1,I_2 \ldots I_n\}$; D$\in\{D_1,D_2 \ldots D_n\}$; F$\in\{F_1,F_2 \ldots F_n\}$; I=Intensity; D=Depth; F=Optical Flow.

We consider that the disadvantage of CoCML is that it requires using an identical CNN model. This weakness is a considerable restriction if distinct modalities improve the learning process with a specific CNN architecture and/or with various settings (i.e., learning rate policies and learning algorithms).

E. Incremental Cross-Modality Learning

Incremental cross-modality learning (InCML) consists of sequential learning each image modality to feed a single CNN based on a transfer learning approach. The fundamental principle of transfer learning is that the initial model acquires knowledge about specific data and then reuses that knowledge in another function. Transfer learning carries the weight information from a previous CNN model to a new CNN model which will be trained next [24], [25].

In our incremental cross-modality approach, a first CNN is learned (trained and validated) with the first image modality frames, then a second one, initialized by transfer learning on the primary CNN, is learned on the second image modality frames, and finally a third one initialized on the second CNN, is learned on the last modality image frames. It does not require correlated modality frames nor equal numbers of items (see Fig.3).

This method has some advantages compared with classical methods. One of the benefits is that it is more flexible than the previous cross-modality learning methods. This method allows different settings to be adapted for each classifier (i.e. different learning rate policies and learning algorithms) which leads to better learning for the final classification system. Transferring the weight information from one classifier which was already learned to another one which will be learnt next, increases the ability of the model to discriminate with a distinct point of view for the same standard target class of modalities (i.e. pedestrians or non pedestrians). It allows additional learning with other modality images without changing the concept target class.

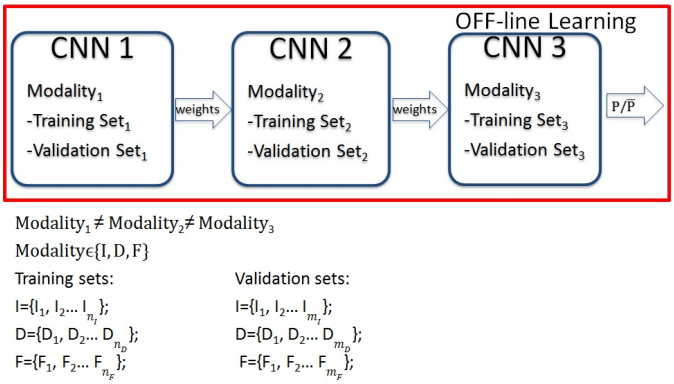


Fig. 3. The Incremental Cross-Modality Learning. The first CNN is learning (training + validation) on the same image modality. When the learning process is over, the weights information from the previous CNN is transferred to the next CNN in which the learning process starts with a different image modality.

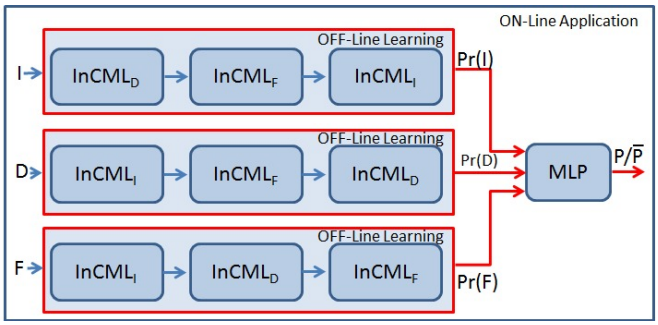


Fig. 4. The Late Fusion Architecture with the Incremental Cross-Modality Learning. The modality probabilistic output scores of Intensity (Pr(I)), Depth (Pr(D)) and Optical Flow (Pr(F)).

Learning this model does not require any calibration and/or synchronization between modality images. This approach could be adapted and utilized when the multi-modality images are not derived from the same database and/ or obtained from related sensors/cameras. Moreover, this procedure can be suitable for using various data sets and stretch out in cross-dataset training.

This approach casts doubt upon whether the learning image modality order could affect the performance of the final classifier. We have examined various combinations by interchanging the imaging modalities, and conclude that to classify the Intensity image modality, the training process needs to start with Depth modality, followed by Optical Flow and finally Intensity images (D, F, I training model of I). The optimal learning order for Optical Flow image modality classification is Depth images, followed by Intensity images and finally Flow images (D, I, F training model of F). To achieve the best learning performance for Depth modality, the training process should start with Intensity images followed by Flow images and finally Depth images (I, F, D for the training of D).

F. Late Fusion Pedestrian Classification with Incremental Cross-Modality Learning

Our late-fusion architecture (see Fig 4) consists of three independent CNN classifiers and an MLP which discriminates between pedestrians (P) and non-pedestrians ($\overline{P}$) based on class probabilistic estimates provided by each CNN. The learning process for each CNN classifier is done with an incremental cross-modality learning approach in an independent manner. The last layer of each CNN provides the modality probability output scores of Intensity Pr(I), Depth Pr(D) and Optical Flow Pr(F). The MLP is composed of three neurons in the input layer, one hidden layer with 100 neurons, and 2 neurons in the output layer. We used the ReLU function for the activation function and a Stochastic Gradient Descent (SGD) [4] solver for the weight optimization. For the weight updates, we used a constant learning rate (1e-07).

Late fusion focuses on three independent components for the learning of modalities, and then, the probabilistic output scores are fused into a multi-modal representation for the final learning step. The off-line learning of the late fusion scheme is therefore costly but it is an efficient solution for on-line applications.

We experimented out target classification task with different CNNs (LeNet, AlexNet, GoogleNet and VGG) and various hyperparameters. Concerning the input image size, the bounding box in Daimler sets are images of 48x96 pixels. We have resized the input layer size accordingly to 48x96 pixels for the LeNet, AlexNet, GoogleNet, VGG. The AlexNet and VGG did not return suitable results for this input image size. To solve this problem the solution is to augment the image input size by interpolation and the best results were obtained with 256x256 pixels for AlexNet, GoogleNet and VGG. However, the performances obtained with those more sophisticated architectures are less than those obtained with LeNet and LeNet+ on the Daimler dataset. Another solution to avoid this problem is to remove some layers and/or to change the convolution parameters and num-output options in the convolution and inner product layers. This is equivalent to designing a task specific CNN architecture.

Each modality CNN, was first set up on the LeNet architecture [20]. We observed that the LeNet has a limited generalization power for our needs. In order to enhance the classification performance and avoid overfitting, we designed a CNN, which we called LeNet+, (see Fig 5) by extending the LeNet architecture by adding three layers and replacing the weight filler algorithm from FC layers. We add a ReLU layer, a Local Response Normalization (LRN) layer [19] at the first Pooling Layer, a Dropout layer [32] with a rate of 50% at the first FC layer. Moreover, for the weight filler we use the Gaussian [21] instead of the Xavier algorithm [14]. For the FC layers, we used 4096 neurons for the first FC layer and two neurons for the second FC layer.

In the next section, we present our set of experiments with CNN-based cross-modality learning for the pedestrian recognition task. We describe the experimental setup and assess the performance of our approaches.

V. Experiments and Results

A. Experimental Setup

The experiments were performed on Daimler stereo vision images of 48 x 96 px with a 12-pixel border around the

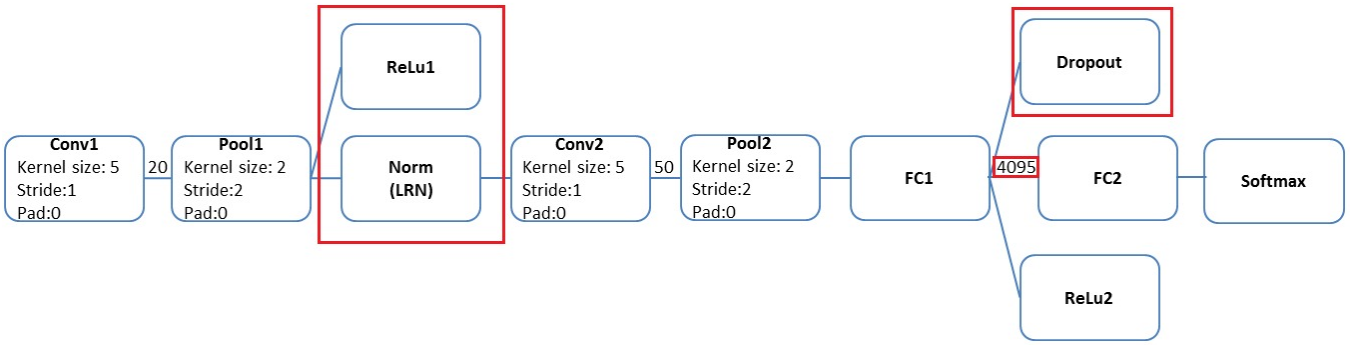


Fig. 5. The proposed extended LeNet Architecture (LeNet+). The extension consists in adding a ReLu and an LRN layer at the first Pooling layer, adding a Dropout layer at the first FC layer, using the Gaussian instead of Xavier algorithm for the weight filler and increasing the outputs for the first FC layer from 500 to 4096.

pedestrian images acquired from three modalities: Intensity, Depth, and Optical Flow.

The learning process (training and validation) was performed on 84577 samples (52112 samples of non-occluded pedestrians and 32465 samples of non-pedestrians), based mainly on the holdout validation method involving a single run. The holdout validation method consists in using a part of the learning set as a validation set (75% samples for training, and the remaining 25% samples for validation) to fit the CNN's hyperparameters. The holdout validation is applied in two steps. In the first step, all the hyperparameters: the learning function, the learning rate policy, the initial learning rate and the number of epochs/iterations are optimized for each image modality among Intensity, Depth and respectively Optical Flow. In the second step, several hyperparameters (the learning function and the learning rate policy) have been fixed to their optimum values obtained in the first step, while only the most critical ones: the initial learning rate and the number of epochs/iterations have been optimized/validated

We also used the 10-fold cross-validation (CV) for fine-tuning those most critical hyper-parameters. The k-fold CV consists of randomly partitioning the training data set into k=10 equal sized subsamples, then a single one is used to validate the model, and the remaining subsamples are used for its training. Since this CV method is time costly only the most critical hyper-parameters (the initial learning rates and the number of epochs/iterations) of the most promising multi-modal InCML classifiers have been optimized.

The testing dataset used to assess the classification performance is independent of the training/validation datasets, and it has the same samples as suggested in the Daimler datasets. It contains 36768 samples of pedestrians (25608 samples of non-occluded pedestrians and 11160 samples of partially occluded pedestrians), and 16235 samples of non-pedestrians.

The learning (training and validation) process for all the proposed CNN models was done with the Caffe Deep Neural Network Framework [16]. The performances are assessed by the Accuracy (ACC) and the Receiver Operating Characteristics (ROC) curves. The complexity of the classification task is evaluated by the area under the curve (AUC). Those performance measures are completed with the F-measure to provide the harmonic average of the precision and recall, which is essential for the object detection system design. The ACC, AUC, F-measure values and ROC curves were executed with the Scikit-Learn tool [26]. We calculate the Confidence Interval (CI) with a confidence level of 95% to evaluate whether one model is statistically better than another one. If the CI of two classifiers are disjoint then the one which is statistically better then other can be chosen.

TABLE I
COMPARISON OF CLASSICAL UNI-MODAL LEARNING (UML) VS PARTICULAR CROSS-MODALITY LEARNING (PACML) ON NON OCCLUDED PEDESTRIAN DATE SET WITH RMSPROP AND POLY LEARNING SETTINGS

	Train on	Valid on	Test on	ACC $\pm$ CI/2
UML	Intensity	Intensity	Intensity	$96.550\,\% \pm 0.174\,\%$
	Depth	Depth	Depth	$89.100\,\% \pm 0.298\,\%$
	Flow	Flow	Flow	$85.690\,\% \pm 0.335\,\%$
PaCML	Depth	Depth	Intensity	$50.510\,\% \pm 0.479\,\%$
	Flow	Flow	Intensity	$73.790\,\% \pm 0.421\,\%$
	Intensity	Intensity	Depth	$58.240\,\% \pm 0.472\,\%$
	Flow	Flow	Depth	$54.230\,\% \pm 0.477\,\%$
	Intensity	Intensity	Flow	$72.970\,\% \pm 0.425\,\%$
	Depth	Depth	Flow	$57.550\,\% \pm 0.473\,\%$

$$CI = 2 * 1.96\sqrt{\frac{P(100-P)}{N}}\%. \quad (1)$$

In this formulation, P represents the performance of the classification system (e.g., ACC, AUC) computed from the confusion matrix, and N represents the number of testing samples. We also measured the Structural Similarity (SSI) [39] and computed the Correlation Coefficient (R) between two couples of images (Intensity-Flow, Intensity-Depth, Depth-Flow) to better analyze the classifier performances and to estimate the area of interest for the proposed cross modality-learning methods.

B. Evaluation of Uni-Modal Learning Classifiers

In order to test the CNN's performance, we carried out several experiments. In our first experiment [27], we investigated the performances of AlexNet and LeNet on the Caltech dataset where pedestrian bounding boxes (BBs) are more than 50 px. All BB were resized to quadratic size (64 x 64 px).

TABLE II
MEAN OF THE STRUCTURAL SIMILARITY INDEX (MSSI) ON THE ORIGINAL IMAGES

MSSI	Intensity Depth	Intensity Flow	Depth Flow
Pedestrians Train Sets	0.1430	0.1592	0.3319
Non Pedestrian Train Sets	0.1150	0.1399	0.3213
Non-Occluded Pedestrians Test Sets	0.1335	0.1529	0.3058
Non Pedestrian Test Sets	0.1129	0.1446	0.2865

TABLE III
MEAN OF CORRELATION COEFFICIENT (MR) ON THE ORIGINAL IMAGES

|MR|	Intensity Depth	Intensity Flow	Depth Flow
Pedestrian Train Sets	0.0011	0.0117	0.0433
Non Pedestrian Train Sets	0.0575	0.0358	0.1222
Non-Occluded Pedestrians Test Sets	0.0359	0.0077	0.0402
Non Pedestrian Test Sets	0.0170	0.0252	0.0752

We observed that the LeNet provides the best results for those small size image datasets.

In the second experiment [28], we evaluated the LeNet architecture with various learning algorithms: Stochastic Gradient Descent (SGD) [4], Adaptive Gradient [8], RMSPROP [35], ADADELTA and learning rate policies: Fixed (FIX), Exponential (EXP) [33], Step Down, Polynomial Decay (POLY) [3], Sigmoid, Multi-Step and Inverse Decay. Each modality classifier was exclusively trained with images of its own modality. We used a fixed batch size of 64 images which means that the training set (63433 samples) needs 992 iterations for one epoch. The holdout validation provides the optimal hyper-parameters for the Intensity modality: 29760 iterations and 0.01 initial learning rate using the RMSPROP learning algorithm (RMS-decay, τ=0.98) and POLY (power, ρ=0.75) learning rate policy; for the Depth modality: 29760 iterations and 0.01 initial learning rate using SGD learning algorithm (gamma, γ =0.99; momentum, μ=0.89) and EXP learning rate policy; for the Optical Flow modality: 29760 iterations and 0.01 initial learning rate using ADADELTA learning algorithm (momentum, μ=0.89) and FIX learning rate policy. We conclude that various modalities require different learning algorithms and rate policies for efficient learning but an equivalent number of iteration and similar initial learning rate. We obtained ACC=96.55% on the Intensity modality (see Table I) followed by the Depth modality with ACC = 89.78% and finally the ACC = 87.34% for Optical Flow.

TABLE IV
MEAN OF THE CORRELATION COEFFICIENT (MR-LOG) ON THE EDGE DETECTOR IMAGES (USING THE LAPLACIAN OF GAUSSIAN METHOD)

MR-LOG	Intensity Depth	Intensity Flow	Depth Flow
Pedestrians Train Sets	0.0126	0.0106	0.0178
Non Pedestrians Train Sets	0.0128	0.0111	0.0253
Non-Occluded Pedestrians Test Sets	0.0142	0.0085	0.0149
Non Pedestrians Test Sets	0.0154	0.0139	0.0198

C. Evaluation of the Particular Cross-Modality Learning classifier

We tested the particular cross-modality learning (PaCML) models where each CNN-based classifier is learnt on one modality with the holdout validation method but tested on a different one (see Table I). The best performance for this approach is achieved on Intensity images when trained on Flow images (ACC = 73.79%), on Depth images when trained on Intensity images (ACC = 58.24%), and on Flow images when trained on Intensity images (ACC = 72.97%). The performances are below those obtained when the learning and testing are performed on the same modality. This idea could be a promising one for automatic annotation of modality images with a classifier learnt with other modality data.

In order to estimate the generalization skills of the proposed automatic annotation approach, we need to know whether this ability depends on the similarity and/or correlation between two modalities. Therefore we compute the mean of the Structural Similarity Index (MSSI) (see Table II) and the mean of the Correlation Coefficient (MR) (see Table III) on the original images and on the edge detector images (using the Laplacian of the Gaussian method) (see Table IV) between a pair of images among Intensity-Flow, Intensity-Depth, and Depth-Flow on the training and testing sets.

As reported, (see Tables II, III, IV) the Depth with Optical Flow is the most correlated modality pair for MSSI similarity, MR and MR-Log correlation coefficients for all investigated data sets. However, even the highest MSSI similarity between Depth and Optical Flow in the original images is of MSSI=0.3319 which proves a low correlation. This highlights the generalization capability of proposed the PaCML model.

Nonetheless, the best performance was obtained with the following particular cross-modality models: learnt on Intensity and tested on Flow and respectively learnt on Flow and tested on Intensity. This method raises the question of whether we can regenerate data in one domain by the observation from the other domain. The Depth modality could not be regenerated only from the Intensity modality because two stereo images are needed (space redundancy). The Flow modality could be created from intensity modality if one has access to images from previous times (temporal redundancy).

D. Comparison of Uni-Modal Classifiers with Cross-Modality Learning Models

In this section, all the models were learnt on the LeNet architecture with the same settings (the optimal ones found previously for the most performant Intensity modality) for the learning algorithm (RMSPROP -RMS-decay, τ=0.98) and for the learning rate policy (POLY power, ρ=0.75), and tested on the non-occluded pedestrian Daimler dataset. The CNNs were enhanced with holdout validation method on the learning set through an optimal number of iterations (29760), and an optimal initial learning rate (0.01) for classical uni-modal learning method and all cross-modality learning models except for the correlated cross-modality one. Since the complexity of the CNNs learning algorithm for the correlated cross-modality learning was extended, the holdout validation provided an

TABLE V
COMPARISON OF CORRELATED (COCML), SEPARATE (SECML) VS INCREMENTAL CROSS-MODALITY (INCML) LEARNING MODELS ON THE NON-OCCLUDED DAIMLER PEDESTRIAN DATA SET. THE RESULTS IN BOLD ARE STATISTICALLY BETTER THAN THOSE OBTAINED WITH THE CLASSICAL UNI-MODAL METHOD.

CNN	Learning Settings	Validation Method	Approach	Train on	Valid on	Test on	ACC ± CI/2
LeNet	Same Settings for RMSPROP with POLY	Holdout	Classical Uni-modal	Intensity	Intensity	Intensity	$96.550\,\% \pm 0.174\,\%$
				Depth	Depth	Depth	$89.100\,\% \pm 0.298\,\%$
				Flow	Flow	Flow	$85.690\,\% \pm 0.335\,\%$
			SeCML	Intensity	Depth	Intensity	$96.310\,\% \pm 0.180\,\%$
				Intensity	Flow	Intensity	$96.230\,\% \pm 0.182\,\%$
				Depth	Intensity	Depth	$89.000\,\% \pm 0.299\,\%$
				Depth	Flow	Depth	$89.330\,\% \pm 0.338\,\%$
				Flow	Intensity	Flow	$86.120\,\% \pm 0.331\,\%$
				Flow	Depth	Flow	**86.600% ± 0.325%**
			CoCML	Intensity_i+Depth_i+Flow_i i=$\overline{1,n}$	Intensity_j+Depth_j+Flow_j j=$\overline{1,m}$	Intensity	$94.540\,\% \pm 0.217\,\%$
						Depth	$85.390\,\% \pm 0.338\,\%$
						Flow	**88.26% ± 0.308%**
					Intensity	Intensity	$94.400\,\% \pm 0.220\,\%$
					Depth	Depth	$86.060\,\% \pm 0.331\,\%$
					Flow	Flow	**87.38% ± 0.318%**
			InCML	Depth_i,Flow_i,Intensity_i i=$\overline{1,n}$	Depth_j,Flow_j,Intensity_j j=$\overline{1,m}$	Intensity	$96.700\,\% \pm 0.171\,\%$
				Intensity_i,Flow_i,Depth_i i=$\overline{1,n}$	Intensity_j,Flow_j,Depth_j j=$\overline{1,m}$	Depth	$89.390\,\% \pm 0.295\,\%$
				Intensity_i+Depth_i+Flow_i i=$\overline{1,n}$	Intensity_j+Depth_j+Flow_j j=$\overline{1,m}$	Flow	**87.02% ± 0.3220%**
	Optimal Specific Settings	K-fold Cross Validation K=10	InCML	Depth_i,Flow_i,Intensity_i i=$\overline{1,n}$	Depth_j,Flow_j,Intensity_j j=$\overline{1,m}$	Intensity	**97.50%± 0.149%**
				Intensity_i,Flow_i,Depth_i i=$\overline{1,n}$	Intensity_j,Flow_j,Depth_j j=$\overline{1,m}$	Depth	88.92% ± 0.300%
				Intensity_i+Depth_i+Flow_i i=$\overline{1,n}$	Intensity_j+Depth_j+Flow_j j=$\overline{1,m}$	Flow	**88.70% ± 0.303%**
LeNet+	Optimal Specific Settings	K-fold Cross Validation K=10	InCML	Depth_i,Flow_i,Intensity_i i=$\overline{1,n}$	Depth_j,Flow_j,Intensity_j j=$\overline{1,m}$	Intensity	**97.78% ± 0.141%**
				Intensity_i,Flow_i,Depth_i i=$\overline{1,n}$	Intensity_j,Flow_j,Depth_j j=$\overline{1,m}$	Depth	**91.30% ± 0.27%**
				Intensity_i+Depth_i+Flow_i i=$\overline{1,n}$	Intensity_j+Depth_j+Flow_j j=$\overline{1,m}$	Flow	**89.75% ± 0.29%**

TABLE VI
OPTIMAL LEARNING RATE AND NUMBER OF ITERATIONS FOR THE INCREMENTAL CROSS MODALITY LEARNING WITH K=10 CROSS-VALIDATION FOR LENET AND LENET+ ARCHITECTURES

Image modality	CNN	Initial Learning Rate		Iterations
		Specific	Averaged	
Intensity	LeNet	0.01	1.5e-05	158640
	LeNet+	0.001	1.2e-05	119040
Depth	LeNet	0.01	1.93e-04	208320
	LeNet+	0.001	1.014e-05	208320
Optical Flow	LeNet	0.01	1.5e-04	158640
	LeNet+	0.01	1.2e-05	158640

optimal number of training iterations augmented to 89220 for an initial learning rate (0.01).

1) Separate Cross-Modality Learning Approach: We evaluated the separate cross-modality learning models where each CNN-based classifier is trained and tested on one image modality but validated (holdout validation method) on a different one. These experiments prove that the cross-modality learning approach performs slightly better than the classical learning approach (see Table V), but only for the Optical Flow and Depth modalities. The improvements are statistically significant only for Optical Flow ΔACC=0.25% (validated in Depth). This could be explained by the fact that for the Depth-Flow modality pair, the values of the MSSI, MR, MR-LOG (see Tables II, III, IV) are stronger than for the other modality pairs (Intensity-Depth, and Intensity-Flow).

2) Correlated Cross-Modality Learning: Since the RMSPROP with POLY learning rate settings produced successful results on the Intensity modality, we used those learning settings for all correlated cross-modality (CoCML) models.

The CoCML models are validated following two different approaches on the multi-modal union data set or on a uni-modal dataset (see Table V). The multi-modality union validation approach yields better results than the uni-modal validation approach. This method performs better than classical uni-modal learning, but only on the Optical Flow testing set, the improvement being statistically significant at ΔACC=1.927%. The experiment could explain this problem, with vast (three times more) and different modalities (Intensity, Depth, and Optical Flow) training data, the breadth and depth of the network should be extended. Moreover, according to [20], the complexity would be limited by the computing resources, which would thus hinder the performance.

3) Incremental Cross-Modality Learning: Since the incremental cross-modality learning (InCML) method is the most

TABLE VII
THE PERFORMANCE WITH LATE FUSION ON NON-OCCLUDED PEDESTRIAN DAIMLER TESTING SET. THE RESULTS IN BOLD ARE STATISTICALLY BETTER THAN THOSE OBTAINED WITH CLASSICAL UNI-MODAL LEARNING. SM=SAME SETTINGS, SP=SPECIFIC SETTINGS, K-CROSS=K-FOLD CROSS-VALIDATION.

CNN	Late-fusion	Trained on	AUC ± CI/2	ACC ± CI/2	F1-Measure ± CI/2
LeNet	Classical Learning	SM	$97.040\,\% \pm 0.162\,\%$	$97.460\,\% \pm 0.150\,\%$	$97.3100\,\% \pm 0.1609\,\%$
LeNet+		SM	$97.560\,\% \pm 0.153\,\%$	$97.970\,\% \pm 0.140\,\%$	$97.4600\,\% \pm 0.1565\,\%$
LeNet	Incremental Cross Modality Learning	SM	$97.200\,\% \pm 0.158\,\%$	$97.620\,\% \pm 0.146\,\%$	$97.4900\,\% \pm 0.1556\,\%$
		SP	$97.47\,\% \pm 0.15\,\%$	$97.690\,\% \pm 0.143\,\%$	$97.5400\,\% \pm 0.1540\,\%$
		SP; K-Cross	*98.26% ± 0.125%*	*98.29%± 0.124%*	*98.60% ± 0.1168%*
LeNet+	Incremental Cross Modality learning	SP; K-Cross	***98.811% ± 0.1039%***	***98.817% ± 0.1036%***	**99.11% ± 0.0934%**

TABLE VIII
THE PERFORMANCE WITH LATE FUSION ON PARTIALLY OCCLUDED PEDESTRIAN DAIMLER TESTING SET. THE RESULTS IN BOLD ARE STATISTICALLY BETTER THAN THOSE OBTAINED WITH CLASSICAL UNI-MODAL LEARNING. SM=SAME SETTINGS, SP=SPECIFIC SETTINGS, K-CROSS=K-FOLD CROSS-VALIDATION.

CNN	Late-fusion	Trained on	AUC ± CI/2	ACC ± CI/2	F1-Measure ± CI/2
LeNet	Classical Learning	SM	$78.130\,\% \pm 0.489\,\%$	$81.110\,\% \pm 0.463\,\%$	$80.6600\,\% \pm 0.4677\,\%$
LeNet+		SM	$84.930\,\% \pm 0.423\,\%$	$82.490\,\% \pm 0.450\,\%$	$82.4800\,\% \pm 0.4502\,\%$
LeNet	Incremental Cross Modality Learning	SM	$78.360\,\% \pm 0.487\,\%$	$80.480\,\% \pm 0.469\,\%$	$79.5700\,\% \pm 0.4775\,\%$
		SP	$78.400\,\% \pm 0.535\,\%$	$81.300\,\% \pm 0.461\,\%$	$80.8700\,\% \pm 0.4658\,\%$
		SP; K-Cross	*82.88% ± 0.446%*	*85.09% ± 0.421%*	*84.65% ± 0.4269%*
LeNet+	Incremental Cross Modality Learning	SP; K-Cross	***86.12% ± 0.409%***	***88.38% ± 0.379%***	**88.34% ± 0.3801%**

promising approach, we decided to carry out more extensive experiments. Thus, the InCML models were learnt using different approaches:

a) Training and holdout validation using the same settings for the learning algorithm (RMSprop with RMS-decay, τ=0.98), for the learning rate policy (POLY with power, ρ=0.75) and a batch size=64 for all three modality-specific CNNs;
b) Training and holdout validation using optimal modality-specific hyper-parameter settings for each CNN. For the Intensity modality: RMSPROP with RMS-decay, τ=0.98 and POLY with power, ρ=0.75; for the Depth modality: SGD with gamma, γ=0.99; momentum, μ=0.89 and EXP; for the Optical Flow modality: ADADELTA with momentum, μ=0.89 and FIX learning rate policy (see Section V.B);
c) Training and k-fold cross-validation method using the algorithm settings from point (a);
d) Training and k-fold cross-validation method using the algorithm settings from point (b);

The holdout validation in (a) and (b) approaches makes it possible not only to fit the optimal initial learning rate, but also to verify that 29760 iterations avoid under and over fitting. The k-fold cross-validation in (c) and (d) approaches has started learning with specific initial learning rates for each modality CNN based on LeNet and respectively LeNet+ architecture for all ten train/valid folds. For each fold and modality CNN we considered the final learning rate for 29760 iterations. The optimal initial learning rate value for each modality CNN are obtained by averaging the final values from prior training folds. These optimal values are used to initialize the train of each modality CNN in a holdout validation method. This makes it possible to find out the optimal number of iterations for the last CNN within each InCML model. The optimal hyperparameters values used in the last learning process are depicted in Table VI.

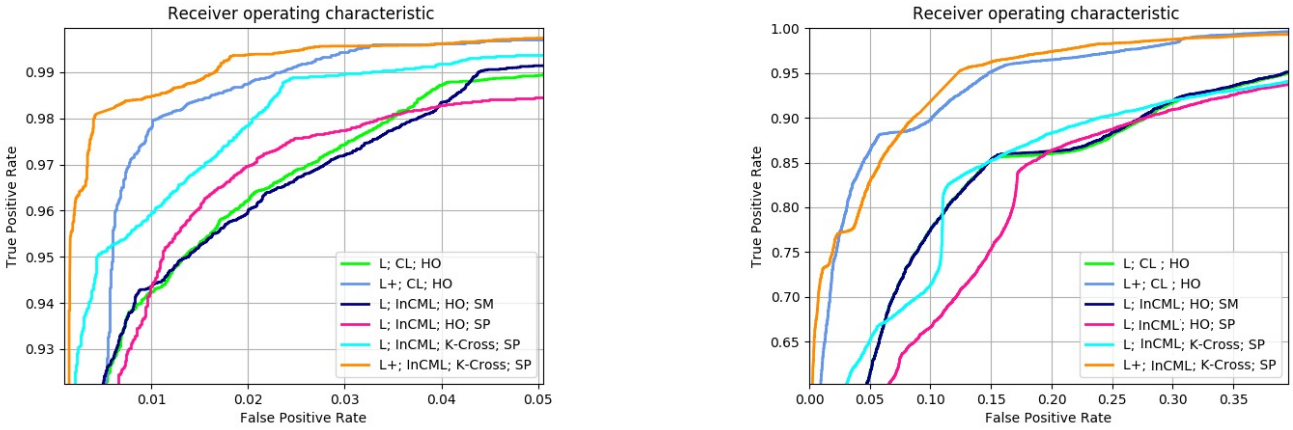
As shown in Table V, the InCML learning approach based on the LeNet architecture with the holdout validation method and RMSPROP - POLY settings, performs slightly better than classical uni-modal learning for all image modalities, but the improvements are statistically significant only for the Optical Flow modality. The LeNet+ architecture we have proposed, with the K-fold cross-validation method and optimal specific learning settings, performs better than the classical learning approach, for all image modalities and the improvements are statistically significant for all image modalities: ΔACC_I =0.915%, ΔACC_D=1.632%, ΔACC_F=3.435% (see Table V). Moreover, this approach is more flexible, allowing for adaptive settings according to each CNN classifier whereas the correlated cross-modality method requires using a single CNN model and therefore the same learning settings.

E. Late-fusion with Classical vs Cross-Modality Learning

The late fusion approach was investigated in two ways with the LeNet and the LeNet+ architecture we proposed, both on the non-occluded pedestrian and the partially occluded pedestrian data sets.

First, all the models were learnt using the same settings with the incremental cross-modality approach. We chose the RMSPROP with the POLY settings since they allowed for the best results on the Intensity modality. Then, optimal specific parameters (selected from the validation set) were used to learn the CNNs through the incremental cross-modality model, and consequently the RMSPROP with the POLY settings for the Intensity modality, the SGD with EXP settings for the Depth modality and ADADELTA with FIX settings for the Flow modality.

The LeNet+ architecture performs statistically better than LeNet for both non-occluded (see VII) and partial-occluded



(a) Non Occluded Pedestrian Data Classification

(b) Partially Occluded Pedestrian Data Classification

Fig. 6. The ROC classification performance on Daimler testing data set; where L=LeNet Architecture, L+= Extended LeNet Architecture, CL=Classical Learning method, HO=Holdout validation, SM=Same Settings, SP=Specific Settings, K-Cross=K-fold Cross-Validation.

(see VIII) Daimler data set not only with classical learning but also for the incremental cross-modality learning. Indeed the confidence intervals are disjoint.

In Tables VII and VIII we show that the performance obtained with incremental cross-modality using the best specific modality learning settings are statistically better than those obtained with the same learning settings. The incremental cross-modality learning is an efficient solution not only with the single modality classifiers, but also with the late-fusion scheme, since its performance is statistically better than late fusion with classical learning. The improvements for non-occluded pedestrian and partially occluded pedestrian Daimler datasets are respectively: $\Delta ACC_{non-occluded} = 1.1034\%$, $\Delta AUC_{non-occluded} = 1.5047\%$, $\Delta ACC_{partially-occluded} = 5.281\%$, $\Delta AUC_{partially-occluded} = 5.497\%$. The improvements are higher for partially-occluded pedestrian recognition than for non-occluded pedestrian recognition. This result proves the robustness of our models. These assessments are also drawn in the ROC curves (see Fig 6). We observe that the ROC curves obtained with the InCML based models with K-Cross and specific settings (SP) are statistically better than all the others approaches but the improvement obtained with LeNet+ vs LeNet is limited.

F. Comparisons with the state-of-the-art methods

We choose to compare our best classifier LeNet+ with Incremental Cross-Modality learning with specific learning settings and the K-fold Cross Validation method (L+; InCML; K-Cross; SP) with the state-of-the-art classifiers provided on the Daimler data sets. Those classifiers are based on a mixture of experts (MoE) with handcrafted features HOG/linSVM [10] and respectively HOG+LBP/MLP [11] within a late fusion of Intensity, Depth and Optical Flow modalities. We also considered for comparison the best Deep models provided on the Daimler dataset, based on Deformation Part and Boltzmann Machine (Deep DP-BM) [22].

For the comparison, we cannot draw the ROC curves of these classifiers since the algorithms source codes is not provided, nor is a detailed explanation of the learning methodology given. Thus, no information is given concerning the learning settings for MLP (e.g., learning rate, number of iterations), nor for SVM (e.g., penalty parameter C of the error term, tolerance for stopping criteria, loss function) or how those hyper-parameters were optimized. Since we do not know how the learning set was distributed between the training and validation sets and whether a cross-validation or a holdout validation technique was used, we cannot reproduce the classification method in a fair manner.

TABLE IX
COMPARISON OF OUR MODELS WITH THE STATE-OF-THE-ART WITH THE FALSE POSITIVE RATE AT 95% TRUE POSITIVE RATE ON DAIMLER DATA SET

Method	Pedestrian Data Set	FP Rate ± CI/2
Deep DP-BM [22]	Partially Occluded	0.25 ± 0.0043%
HOG/linSVM MoE [10]		0.20 ± 0.0040%
L+; InCML; K-Cross; SP		**0.124 ± 0.0033%**
Deep DP-BM [22]	Non Occluded	0.05 ± 0.0021%
HOG/linSVM MoE [10]		0.0302 ± 0.0016%
HOG+LBP/MLP MoE [11]		0.0035 ± 0.00056%
L+; InCML; K-Cross; SP		**0.0016 ± 0.000382%**

Therefore, to assess the performance of our best classifier (L+; InCML; K-Cross; SP), we compute the false positive rates (see Table IX) using a true positive rate of 95% as a frequent reference point using the interpolation method. This target allows a fair comparison with the cited state-of-the-art pedestrian classifiers on both partially-occluded and non-occluded pedestrian Daimler data sets. We also computed the confidence intervals (CI) with a risk level of 0.05 to allow a significant statistical analysis. Our model outperforms both the handcrafted-features MoE and deep DP-BM models.

The improvements obtained with our classifier (L+; InCML; k-Cross; SP) compared with all these models are statistically significant on both partiallyoccluded and nonoccluded data sets since the confidence intervals are disjoint:
ΔFPR MoE$_{partially-occluded}$=0.0484%,

ΔFPR MoE$_{non-occluded}$=0.0019%,
ΔFPR DP-BM$_{partially-occluded}$=0.126%,
ΔFPR DP-BM$_{non-occluded}$=0.076%.
It is interesting to note that the improvement obtained with our model is more significant for the partial-occluded task for both the handcrafted-features and deep models. However, our model needs to be validated on more extensive datasets and various applications (multiclass road obstacle detection, traffic collision risk assessment).

VI. Conclusion

In this paper, we systematically depicted different cross-modality learning approaches of four methods based on Convolutional Neural Networks for pedestrian recognition: (1) a particular cross-modality learning (PaCML); (2) a separate cross-modality learning (SeCML); (3) a correlated cross-modality learning (CoCML) and (4) an incremental cross-modality learning (InCML). The particular cross-modality learning could be extended for an automatic annotation method of new modality images. The incremental cross-modality learning could be used when there are not enough annotated images in each modality to improve the classification performances. The separate and correlated cross-modalities learning models do not allow for statistically significant improvements since they require the same learning settings for all modality models and, for the second one (CoCML) the same image frame for each modality.

The effectiveness of those methods has been analyzed through various performance measures with statistical coefficients (Confidence Intervals, Correlation Coefficients, Structural Similarity Index). Incremental cross-modality learning based on modality transfer learning is better than both the separate and correlated cross-modality learning models. It also improves the classification performances, in contrast to classical learning of uni-modal CNNs through late-fusion designed on the Daimler data set. We assume that the incremental method is the promising cross-modality learning model. Indeed, this cross-modality learning method is more flexible than the others we analyzed since it could be used with different learning settings adapted for each image modality. In order to improve its performances, we proposed a new CNN architecture called LeNet+ which outperforms the state-of-the-art pedestrian classifier for both non-occluded and partially-occluded pedestrian Daimler data set. However, those cross-modality learning methods have to be validated not only for pedestrian classification, but also for pedestrian unit action recognition, pedestrian detection and tracking.

The enhancements proposed in LeNet+ allow us to validate the cross-validation learning methodology and chose from the proposed models (PaCML, SeCML, CoCML, InCML) the most promising one on a multi-modality classification task on the Daimler dataset. The InCML model could be used not only for an ADAS system but also for a wide variety of learning components with a multi-modality system within complex multi-class classifiers. Currently, we are working with CNNs designed for multi-class detection (SSD, Faster RCNN, R-FCN) on different databases. In addition, we intend to apply the promising InCML model for the classification and detection of other road objects (traffic signs and traffic lights) and road users (vehicles, cyclists).

Acknowledgment

This research was funded by Inria Paris and Normandy Region.

References

[1] Anelia Angelova, Alex Krizhevsky, and Vincent Vanhoucke. Pedestrian detection with a large-field-of-view deep network. In *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*, pages 704–711, 2015.

[2] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, pages 613–627, Cham, 2015. Springer International Publishing.

[3] Alexander Borichev and Yuri Tomilov. Optimal polynomial decay of functions and operator semigroups. *Mathematische Annalen*, 347(2):455–478, Jun 2010.

[4] Léon Bottou. Stochastic gradient descent tricks. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[6] Piotr Dollar, Zhuowen Tu, Pietro Perona, and Serge Belongie. Integral channel features. In *Proc. BMVC*, pages 91.1–91.11, 2009. doi:10.5244/C.23.91.

[7] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, April 2012.

[8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Technical Report UCB/EECS-2010-24, EECS Department, University of California, Berkeley, Mar 2010.

[9] M. Eisenbach, D. Seichter, T. Wengefeld, and H. M. Gross. Cooperative multi-scale convolutional neural networks for person detection. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 267–276, July 2016.

[10] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 990–997, June 2010.

[11] M. Enzweiler and D. M. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *IEEE Transactions on Image Processing*, 20(10):2967–2979, Oct 2011.

[12] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, September 2010.

[13] H. Fukui, T. Yamashita, Y. Yamauchi, H. Fujiyoshi, and H. Murase. Pedestrian detection based on deep convolutional neural network with ensemble inference network. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 223–228, June 2015.

[14] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*, 2010.

[15] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[16] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[17] C. Karaoguz and A. Gepperth. Incremental learning for bootstrapping object classifier models. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1242–1248, Nov 2016.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[20] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[21] R. Matei and P. Ungureanu. A class of gaussian-shaped cnn filter banks. In *2008 11th International Workshop on Cellular Neural Networks and Their Applications*, pages 135–139, July 2008.

[22] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3258–3265, June 2012.

[23] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[24] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[25] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[27] Dănuţ Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, and Abdelaziz Bensrhair. Fusion of stereo vision for pedestrian recognition using convolutional neural networks. *In 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 47–52, April 2017.

[28] Dănuţ Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, and Abdelaziz Bensrhair. Incremental cross-modality deep learning for pedestrian recognition. *In 28th IEEE Intelligent Vehicles Symposium (IV)*, pages 523–528, June 2017.

[29] Dănuţ Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, and Abdelaziz Bensrhair. Pedestrian recognition through different cross-modality deep learning methods. In *2017 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pages 133–138, June 2017.

[30] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *2009 IEEE 12th International Conference on Computer Vision*, pages 24–31, Sept 2009.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[33] Peng Sun. Exponential decay of expansive constants. *Science China Mathematics*, 56(10):2063–2067, Oct 2013.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[35] Tieleman T. and Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. In *COURSERA: Neural Networks for Machine Learning*, 2012.

[36] David Vazquez, Antonio M. Lopez, Javier Marin, Daniel Ponsa, and David Geronimo. Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(4):797–809, 2014.

[37] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

[38] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 509–514, April 2016.

[39] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[40] Xiaogang, Pengxu Wei, Wei Ke, Qixiang Ye, and Jianbin Jiao. Pedestrian detection with deep convolutional neural network. In C.V. Jawahar and Shiguang Shan, editors, *Computer Vision - ACCV 2014 Workshops: Singapore, Singapore, November 1-2, 2014, Revised Selected Papers, Part I*, pages 354–365, Cham, 2015. Springer International Publishing.

Dănuţ Ovidiu Pop received an M.Sc. in computer science from Petru-Maior University, Târgu Mureş, Romania, in 2014. He is currently working toward the Ph.D. degree at INRIA Paris, RITS team, Paris, France, in collaboration with Normandie Univ, INSA Rouen, LITIS Laboratory, Rouen, France and the Babeş-Bolyai Univ, Department of Computer Science, Cluj-Napoca, Romania. His research interests include classification, detection, actions prediction and tracking of road users based on vision, radar, and sensors fusion methods for the intelligent vehicle.

Alexandrina Rogozan is associate professor at LITIS Laboratory at INSA of Rouen, France, since 2000. She obtained a Ph. D. title in 1999 at University of Paris IX, France, with a thesis on "Heterogeneous Data Fusion for Audio-Visual Speech Recognition". Her current research activity is concerned with deep learning, multiple kernels, hybrid models and fusion schemes and the corresponding adaptation methods, for automatic classification and understanding. Her privileged application areas are image and text mining, and particularly the intelligent vehicles.

Abdelaziz Bensrhair graduated with the Master of Science in electrical engineering (1989) and the Ph. D. in Computer science (1992) at the University of Rouen, France. He is currently a Professor in Information Systems Architecture Department, head of Intelligent Transportation Systems Division (2007-2012) and co-director of the Computer Science, Information Processing, and Systems Laboratory (LITIS) of the National Institute of Applied Science Rouen (INSAR) (2002-2016).

Fawzi Nashashibi 51 years, is a senior researcher and has been the Program Manager of RITS Team at INRIA (Paris-Rocquencourt) since 2010. Fawzi Nashashibi has a Masters Degree in Automation, Industrial Engineering and Signal Processing (LAAS/CNRS), a PhD in Robotics from Toulouse University prepared in (LAAS/CNRS) laboratory, and a HDR Diploma (Accreditation to research supervision) from University of Pierre et Marie Curie (Paris 6). His main research topics are in environment perception and multi-sensor fusion, vehicle positioning and environment 3D modeling with main applications in Intelligent Transport Systems and Robotics. He is the author of numerous publications and patents in the field of ITS and ADAS systems. Since 1994 he has also been a lecturer at several universities (Mines ParisTech, Paris 8 Saint-Denis, Leonard de Vinci Univ. -ESILV professor, Telecom Sud Paris, INT Evry, Ecole Centrale dElectronique) in the fields of image and signal processing, 3D perception, 3D infographics, mobile robotics and C++/JAVA programming. IEEE member, he is also member of the ITS Society and the Robotics & Automation Society. He is an Associate Editor of several IEEE international conferences such as IV, ITSC, ICARCV and Journals.